Watching without Seeing: A Tool to Surveil Astronaut Health Outcomes While Maintaining Astronaut Medical Privacy

R.J. Reynolds, L.S. Kanikkannan, M. Van Baalen

Background

- We would like to understand the incidence/mortality rates of various medical conditions in the astronaut cohort.
  - Normally, we'd stratify the cohort and calculate each statistic directly.
- Inherent problems with computing statistics by stratification:
  - Particular demographic combinations (age/sex/calendar year) can lead to loss of anonymity by over-stratification (i.e., "identifiability");
  - This is problematic in that not all users may have NTK access for astronaut medical data;
  - Similarly, when outcomes are rare risk measures for them are not robust or meaningful;
- Our solution:
  - Rather than <u>calculate</u> statistics directly by stratification, we will <u>estimate</u> them from statistical models;
    - The curse of dimensionality
  - As a result, the incidence curves generated by the tool will simultaneously be the <u>best estimate possible</u> from the empirical data, but <u>will not be specific</u> to any individual – no matter how narrow or specific the demographic selection criteria.

Methods

- A Poisson regression (or similar model for count data) can then be fit to these data which can provide estimates of *incidence rates* (for disease outcomes or mortality) by using the person-time (e.g. person-years) as the exposure:
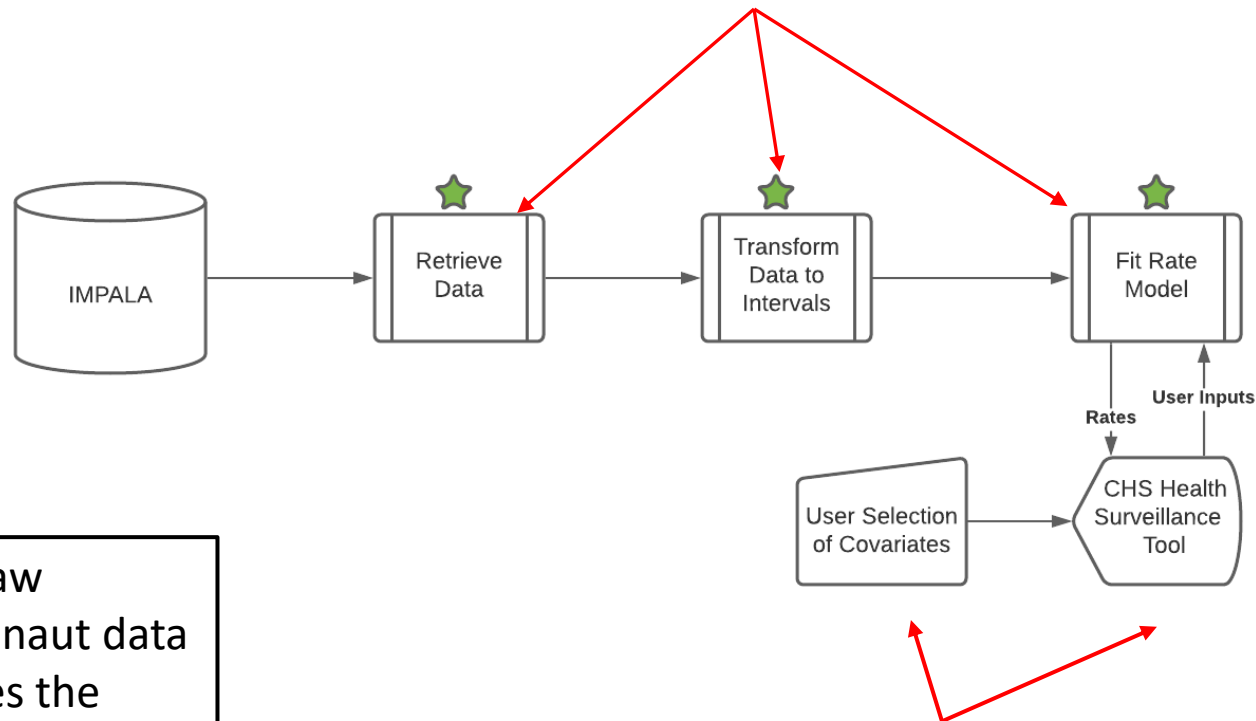
$$\log(\mathrm{E}(Y \mid x)) - \log(\text{exposure}) = \log\left(\frac{\mathrm{E}(Y \mid x)}{\text{exposure}}\right) = \theta' x$$

- For standardized estimates, the *SIR/SMR* can be modeled directly by using the expected counts (from a specific comparison group) as the exposure.

$$Expected_i = (Astronaut\ PYs_i) \times (Comparison\ group\ rate_i)$$

Methods

Automated Python code does these things

IMPALA

Retrieve Data → Transform Data to Intervals → Fit Rate Model

Rates | User Inputs

User Selection of Covariates → CHS Health Surveillance Tool

TOOLS USED:
Hue -> Python (CDSW) ->
Tableau

No raw astronaut data leaves the secure IMPALA environment

End-users operate here, seeing only visual representation of model results, not raw data of any kind

Methods

Dataset construction

DOB, selection date, dates of each flight, date of retirement and DOD are used to form discreet time intervals, tagged with unchanging demographic variables (sex, race/ethnicity, educational and occupational background), time-varying covariates (calendar year, time in space, age) and an interval time width (fractional person-years). Outcomes (diagnoses) and their dates from the IMPALA model data tables can be converted to a binary indicator of an event occurring and the date of the occurrence for integration into the life-period intervals.
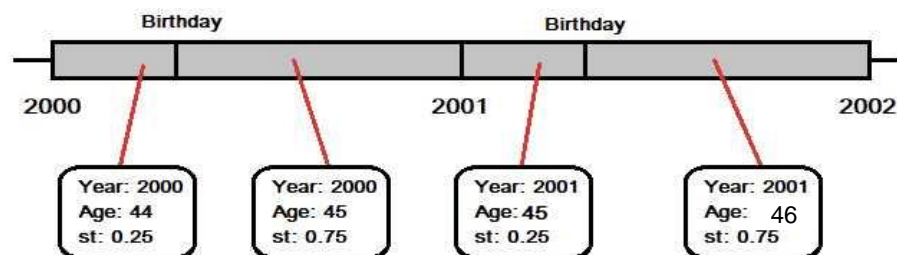
ID	Year	Age	st	outcome
1	2000	44	0.25	0
1	2000	45	0.75	0
1	2001	45	0.25	0
1	2001	46	0.75	0
...				

Results

- The first iteration of the tool computes incidence and mortality rates for cardiovascular conditions and cancers.

- Code has been developed to retrieve the appropriate data from the IMPALA analysis platform, compute the models for incidence and mortality, and then use those models to generate the corresponding rate curves.

- A companion graphical user interface allows the user to specify the curves and visualize the results.

Conclusions

- The rapid surveillance tool described here is neither meant to be a definitive assessment of the incidence or mortality of any particular disease or condition in the astronaut population, nor is it meant to be used for research purposes.

- By automating a repetitive process and leveraging carefully curated astronaut health outcomes, the tool makes possible a rapid "first look" into known areas of concern, and, if used judiciously, may surface new areas of concern for long-term astronaut health.

- Instead, it is explicitly recognized that if a covariate is not statistically significant and not a confounder then it will likely have very little effect on the estimate of the incidence and mortality rates.

- Users are able to specify the disease endpoint of interest and the covariates over which they would like to stratify. The system then uses the resulting model to compute the estimated rates for the user-chosen configuration of variables as visualizes those either over an age range within a specified time-period, or over time for astronauts with a specified age range.